引用格式：崇媛媛, 冯硕, 王嵩, 等. 大模型驱动、人机协同的机器化学家云设施. 中国科学院院刊, 2024, 39(1): 41-49, doi: 10.16418/j.issn.1000-3045.20231205003.
Chong Y Y, Feng S, Wang S, et al. Large model-driven, human-computer collaborative robotic AI-chemist cloud facility. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 41-49, doi: 10.16418/j.issn.1000-3045.20231205003. (in Chinese)

大模型驱动、人机协同的机器化学家云设施

崇媛媛　冯 硕　王 嵩　江 俊*

中国科学技术大学　精准智能化学重点实验室　合肥　230026

摘要　在当前由人工智能引发的科技浪潮中，化学科学研究面临着前所未有的机遇和挑战。为推动化学研究范式的变革，文章提出了机器化学家云设施的建设方案。该系统通过收集多通道数据构建数据库、发展化学知识增强的科学大模型、建设机器人设施集群及搭建智能管理决策系统，践行科研新范式，大幅度提升科研效率，解决终端应用的科学问题。这一基础设施有望推动科研范式变革，在化学领域取得重大科学突破。

关键词　化学研究范式变革，机器化学家云设施，人工智能，自动化实验，人机协作

DOI　10.16418/j.issn.1000-3045.20231205003

CSTR　32128.14.CASbulletin.20231205003

当前，随着大数据与人工智能（AI）技术的飞速发展，人类正迎来新一轮科技与产业革命。一些代表性智能工具，如AlphaFold2和ChatGPT展现出了超越人类解决复杂问题的能力。人工智能技术的引入不仅极大地提升了科研过程中通用工具的效率和准确性，更重要的是，它有助于构建一个由产业需求驱动科学研究的有效体系。本文旨在通过探讨大模型驱动、人机协同的机器化学家云设施建设，进一步探讨“AI for Science”科研新范式的变革。

1 化学研究范式变革：挑战、机遇与趋势

化学作为一门基础科学，致力于研究物质的组成、结构、性质，及其在不同条件下的行为和与其他物质之间的相互作用等。实验和理论两种研究方法相互补充，共同推进化学科学研究的进步，在推动新材料研发、探索新能源、改进生物医疗技术等方面均有

*通信作者
资助项目：国家自然科学基金杰出青年科学基金项目（22025304），国家自然科学基金重点项目（22033007），中国科学院稳定支持基础研究领域青年团队计划（YSBR-005）
修改稿收到日期：2024年1月4日

着重要意义和广泛应用[1-3]。

1.1 化学研究领域面临的挑战

当前，化学科学领域的研究对象日益复杂化和高维度化，这给科学研究带来了巨大的挑战。现行主流的研究方法依赖于穷举试错和降低变量复杂度等传统手段，其低效和局限性问题日益凸显[4]。

从微观的原子、分子尺度到宏观的应用材料尺度，物质性质受不同反应条件及相互作用影响[5-8]，难以进行准确预测和描述。人类对于解析从微观到宏观尺度复杂系统底层规律的渴望日益增长，这种解析将对合成自动优化[9]、按需逆向设计材料[10]，以及精准控制生物医学过程[11]具有重要指导意义。然而，从物理常数、薛定谔方程和元素周期表等基本规则到复杂应用层面的自下而上演化，其中存在极大的复杂度和多样性[12,13]，这就导致现实世界问题与结构-效能关系之间的脱节。化学合成仍然依赖于专家经验，距离智能优化目标相去甚远；数据的不完整和构效关系的不明确，依旧是材料逆向定制设计的“拦路虎”；生物学领域中心法则过程的演变信息缺乏，限制了人类对于疾病机理和生命本质的认知。

1.2 化学研究范式变革的机遇

为了积极应对化学科学领域面临的挑战，必须创新研究方法，革新研究范式。随着大数据时代的来临，数据驱动的科研范式开始崭露头角。人工智能擅长从高维度、高复杂度的数据中探索变量之间的关联[14,15]，这一趋势为迎接挑战提供了全新的机遇。

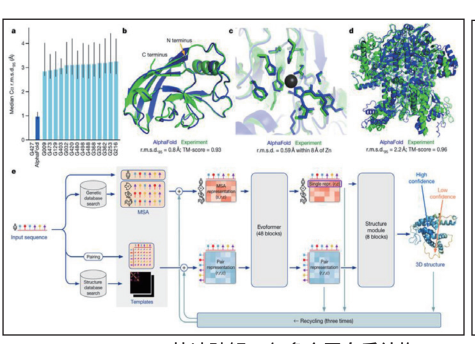
（1）以深度学习和大模型为代表的人工智能技术，具备学习能力、自适应性、自主决策、模式识别与预测等特质，展现出超越人类、智能决策的优势（图1）。2016年，DeepMind公司开发的人工智能程序AlphaGo采用深度强化学习技术，结合深度神经网络和强化学习算法，对棋局策略进行高效搜索并做出精准判断，超越了人类传统的启发式围棋搜索方法[16]。这一围棋界的人机较量成为人工智能发展史上的里程碑事件，首次展现出人工智能在复杂决策领域的应用潜力；2021年，蛋白质结构预测程序AlphaFold2基于深度神经网络和自注意力机制，通过训练大规模序列数据，实现对蛋白质三维结构的高精度预测[17]。这一突破性成果，对药物设计和疾病诊断具有潜在重要意义；2023年火爆全球的对话生成模型ChatGPT，运用Transformer架构中的自注意力机制和多层神经网络生成语言，并通过无监督学习的方式不断迭代语言生成能力，提升人机交互体验①。其代表了人工智能在自然语言处理领域的突破式进展，有望帮助人类获取信息及智能决策，实现通用认知智能的涌现。

（2）智能驱动的机器人化学研究近年来取得了一系列突破。2022年，英国格拉斯哥大学的Cronin团队[18]开发了自动化机器人系统Chemputer，其集成了文献阅读、实验方案定制、化合物合成和表征功能，能够将文献中的合成步骤转换成机器可读的化学描述语言，并储存于内部数据库中，以便机器人自动执行。英国利物浦大学Cooper团队[19]开发了移动机器人化学家，其可以高效执行实验，并使用贝叶斯算法优化驱动，根据已有实验数据进一步分析优化实验计划；不过，Cooper认为目前的机器人缺乏计算大脑，没有利用已有的化学知识，不能引入理论或物理模型，以致贝叶斯优化是盲目的。2022年，中国科学技术大学江俊团队[20]开发出数据智能驱动的全流程机器化学家，其由机器阅读系统、机器计算系统和机器实验系统三大模块组成，能学习前人知识与智慧，思考产生物理模型及提供智能预测，并高效实验产生全生命周期数据。该平台充分发挥机器数据可重复、可信任、可溯源、可对齐的优点，用精准实验数据校准理论的预训练模型，实现了理实交融的智能预测。

① OpenAI. GPT-4 Technical Report 2023. (2023-03-01)[2024-01-02]. https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O.




Alpha-Go 围棋的棋力超过人类职业围棋顶尖水平


Alpha-Fold2 快速破解 2 亿多个蛋白质结构，加速了结构生物学的发展



ChatGPT/GPT-4 实现了通用认知智能

图 1　人工智能展现出解析复杂体系的强大能力

Figure 1　Artificial Intelligence (AI) demonstrates powerful ability to parse complex systems

1.3 化学科学研究的发展趋势

（1）国际情况。自从ChatGPT在2023年初证明了通用认知智能的可行性，不到半年时间，美国、英国、加拿大、荷兰、瑞士等国纷纷加速投入发展装备智能科学大模型的智慧大脑。2023年，美国更新发布《国家人工智能研发战略计划》，每年投入大量经费用于支持数据科学、人工智能、量子信息等研究②；2023年起，英国也投入经费开始用于融合大模型、机器人和智能联盟的智能创新工场建设；2023年4月，加拿大向加速联盟追加15亿元人民币投资用于建设大规模智能实验室基础设施；2023年7月，荷兰开始打造机器人化学实验室；2022年12月，瑞士投入资金用于打造公共服务设施，借助大模型来驱动机器人。具有化学智慧的机器科学家支撑产业数字化，也已经成为现实。2022年，联合利华60%的年度研发经费都用于购买英国利物浦大学机器化学家材料创新工厂提供的智能合成与测试服务。

（2）我国情况。目前，我国在覆盖智能文献调研与研究规划、计算、实验、优化全流程的机器化学家系统方面局部领先，但在大规模智能实验室与化学科学大模型方面亟须进行建制化的项目部署，避免“起个大早，赶个晚集”。

2 机器化学家云设施：未来化学研究新工具

人工智能技术的飞速发展正在为化学科学研究带来前所未有的机遇和挑战。在当前科技革命和产业变革的浪潮中，研发汇聚科学数据、人工智能算法、智能机器人和云平台的化学研究新工具成为迫切而必要的任务。这一工具的研发有望解决长期困扰化学科学革新的维度灾难和复杂巨系统黑盒问题，从而推动我国在高值化学品、功能材料、生物化学医药等领域取得颠覆性突破。

2.1 机器化学家云设施的内涵

人类化学家传统的研究工作路线通常由提出需求、查阅文献、设计方案、理论模拟和实验验证、提炼理论及解决实际问题等流程组成。相较于人类的研究方法论，涵盖数据库、人机交互、机器实验员、化学工作站和化学大脑的机器化学家云设施不仅能够完全覆盖以上流程，还可以通过人机交互系统进行人机协同定制化解决特定难题（图2）。

数据是现代科学研究的重要组成部分，对于机器化学家云设施而言更是至关重要。通过数据库中海量

② National artificial intelligence research and development strategic plan 2023 update. (2023-05-23)[2024-01-02]. https://www.nitrd.gov/national-artificial-intelligence-research-and-development-strategic-plan-2023-update.

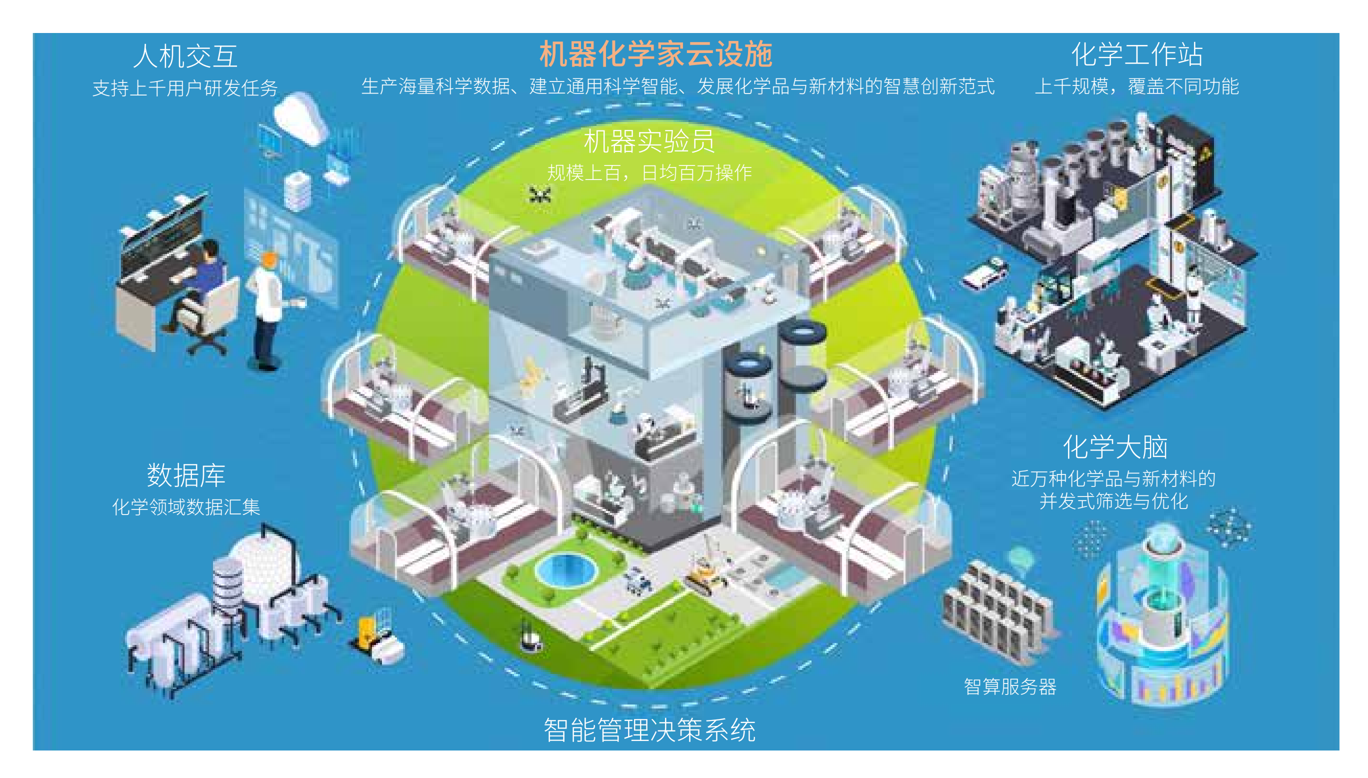


图2　机器化学家云设施组织形态

Figure 2　Framework of robotic AI-chemist cloud facility

化学数据驱动，机器化学家可以学习前人知识与智慧，人机交互提出科学问题，再经过融合科学大模型的化学大脑进行思考，建立物理模型并提供智能预测。随后给出研究方案，驱动高效的机器实验员、化学工作站及智算服务器产生高质量的实验数据和理论模拟数据。再通过数据反馈优化科学大模型，形成垂直领域的应用模型，从而解决具体科学难题。其独特之处在于，能够高效地整合数据知识、不断调整理论和实验设计，实现全流程的智能化推演。目前，中国科学技术大学研制成功了全球首个数据智能驱动的全流程机器化学家，中国科学院自动化研究所与武汉人工智能研究院推出了“紫东太初”全模态大模型，科大讯飞研制了“讯飞星火”认知大模型，中国科学院在全国部署了20余个科学数据中心、智算中心，具备建设机器化学家云设施的良好基础。

机器化学家云设施将带来全新的科研组织形式，即机器实验员实现科研人员体力的解放，数据库和化学大脑实现科研人员脑力的解放，云平台智能管理决策系统实现个体间的链接，并通过新生成数据与科学大模型间的相互对抗校准，逐步进行全局优化，通力合作帮助科研用户取得科学突破。整套设施将大力推动我国科研组织形式的变革，实现大数据、认知智能、机器学习、智能硬件等多领域的高度融合，推动科学研究向更深、更广领域拓展。

2.2 建设机器化学家云设施的意义

机器化学家云设施在实验机器人硬件中融合了深度学习和科学大模型等人工智能技术，为聚合多学科方法论、融合多领域知识逻辑、耦合化学科学家群体智慧、减轻实验人员工作强度提供了技术底座，将加速实验设计和数据分析过程，提升化学科学研究的效率和准确性。当前，美英等多国纷纷加速投入发展装备有科学大模型的机器科研工具。而智能领域是典型的“赢家通吃”，几乎没有后发优势，只有抢占先机，率先掌握先进科研工具才能使我国在新一轮科技革命

中不受制于人。因此，把握我国自主研发机器化学家的领先优势，研制机器化学家云设施，能够防范我国在智能化学研究新范式的基础研究工具方面被“卡脖子”，争抢智能化学领域优势地位。机器化学家云设施的建设也将对整个社会产生积极的溢出效应，推动产业数字化，提升生产效率，还有望催生新一轮的产业革命。

总体而言，机器化学家云设施对增强我国在科技创新领域的竞争力、确保在新兴科技领域中的领先地位具有巨大而深远的意义，将助力我国在全球新一轮科技革命中取得更大的发展和突破。

3 机器化学家云设施：分层架构

通过科学大模型预测和智能机器人实证相互对抗、协同进化，打造具备化学科学智能的机器化学家云设施，将驱动研究范式变革，产生重大科学突破。

3.1 化学科学数据库

数据驱动的研究范式中，科学数据的有效整合和利用是创新的核心驱动力。然而，当前科学数据普遍存在标准不统一、质量良莠不齐、多来源数据相对独立等问题，限制了基于数据的化学科学研究。因此，迫切需要打破数据孤岛现象，融合不同来源的理论和实验数据构建多学科知识和多模态数据的人工智能化学科学数据库。这将为化学科学领域的智能发展提供坚实的数据基础。

化学科学数据库将嵌入人工智能模型，并汇聚文献数据、整合理论与实验数据，包括以下4个方面。

（1）*化学科学领域数据汇聚。*整合各单位数据资源，利用科学文献中的文本、表格、图像等多模态数据，以及第一性原理模拟所产生的大量化学分子和材料的基础物理化学数据。同时，建立实验数据采集渠道和国家标准，实现标准化数据的自动采集和快速分析。

（2）*科技文献机器阅读工具建设。*通过对来源于科技期刊、教科书、题库等语料数据的清理、筛选和标注，获取高价值通用领域预训练语料和化学科学领域预训练语料。利用深度挖掘技术对科技文献内容进行深入挖掘，从文本、图像、表格中提取计算和实验数据。

（3）*数据精编与高质量数据库建设。*标注预训练语料，对文献中的计算和实验数据进行整编，并进行数据分类和质量评估。开发基于可解释模型的数据鉴别和质量评分技术，以智能方式清洗数据。

（4）*知识嵌入与知识图谱构建。*运用映射关系分析构建关联模型，建立化学科学知识图谱，包括结构、性质、演化关联性。通过知识图谱引导多模态数据融合，构建统一、高效、可扩展、结构清晰的数据存储格式。利用预训练模型等工具将知识图谱嵌入到化学科学大模型中，提升知识的利用效率。

3.2 科学大模型

当前，基于神经网络的大模型在预测方面存在可靠性不高、逻辑推理和语义理解深度不足、可解释性和可调试性不强等核心问题，因此在对准确度要求较高的化学科学的应用中表现不佳。针对这些问题，需要发展基于数理逻辑的科学大模型，将数据驱动的神经网络模型与知识驱动的符号逻辑推理引擎深度融合，并应用于数学、化学、物理等智能科学领域。

本研究提出的科学大模型框架在现有数据库、潜在数据库和终端应用基础上，专注于研究知识驱动的推理引擎。该引擎构建在领域本体和知识库之上，并与数据库和潜在数据库连接，以模拟人类思考的认知推理和决策能力，从而弥补大型模型在可靠性、可解释性和可调试性等方面的缺陷。

科学大模型通过综合知识图谱和基于化学认知的知识增强算法，融入专家的化学知识和理解，利用特色化学描述符，创建基于化学原理的清晰人工智能算法，以解决大规模筛选和策略优化等复杂挑战，构建出具备“化学智慧”的机器科学家大脑。根据用户需

求，设计实验方案和运行流程，实时分析实验数据，调整智能模型，并持续反馈优化实验方案，实现实验方案和流程的自动决策与优化。

3.3 机器人平台

机器人平台将提供高效、精准的实验和数据处理解决方案，具体包括以下4个方面。

（1）基于微通道连续流的全自动高通量研究系统。系统旨在精确、自动、高通量地进行重要有机化学反应和关键功能材料合成，需要解决多领域技术问题，并集成多个关键功能子系统，包括多通道反应物自动切换、微通道连续流反应、产物收集和后处理、在线检测和自动采样、色谱接口、反应温度控制、总控和人机交互系统。

（2）全自动高通量研究系统的功能扩展。为确保高通量实验结果可靠性，各子系统配备冗余传感器，并结合视觉识别技术进行实时反馈和异常数据自动筛选。研究人员只需准备反应物库和输入反应矩阵，系统即可完成实验、后处理和检测，并批量输出数据。未来可通过增加子系统和功能模块，拓展研究范围，完成更复杂的后处理和检测分析工作。

（3）全自主实验的移动操作机器人。设计六自由度机械臂与全向移动底盘的软硬件集成；设计实验室环境的视觉感知算法，以及高精度视觉引导、实时力反馈的灵巧控制方法；研究多模态数据的高精度定位和建图方法，开发动态避障算法和任务管理系统，实现移动操作机器人的全自主实验。

（4）全流程智能化学实验室。研制自主知识产权的自动封装机、液体自动分配工作站和电化学自动化测试工作站，同时设计全自主移动操作机器人、高通量实验平台、实验仪器设备的协同控制系统和全流程任务调度系统，以搭建集成化学合成、谱学表征和性能测试等功能的全流程智能化学实验室（图3），实现化学研究的全场景覆盖。

3.4 智能管理决策系统

智能管理决策系统即智能化学云平台，包括机器化学家指令集、操作系统、联邦学习算法系统等，促使机器化学家能够在不同实验任务和实验室之间进行迁移学习，最终构建标准化的云平台层面智能化学实验室（图4）。

指令集包括开发接口函数、通信协议、设备规范和数据标准4个部分的标准化，以支持不同来源的数据对齐，实现数据互联互通。具有友好人机交互界面、明确业务流、直观数据可视功能的操作系统帮助科研人员摆脱物理空间限制，远程即可开展实验、模拟和数据分析，同时也便于系统整体进行实验任务调配、合理安排资源。联邦学习算法系统的核心是数据不动、模型动，即可以在保障数据隐私安全的前提下，支持不同用户不同实验室之间的数据共享。

通过发布智能化学实验室的标准规范，实现在云端共享使用数据库和人工智能模型。这一系统旨在实

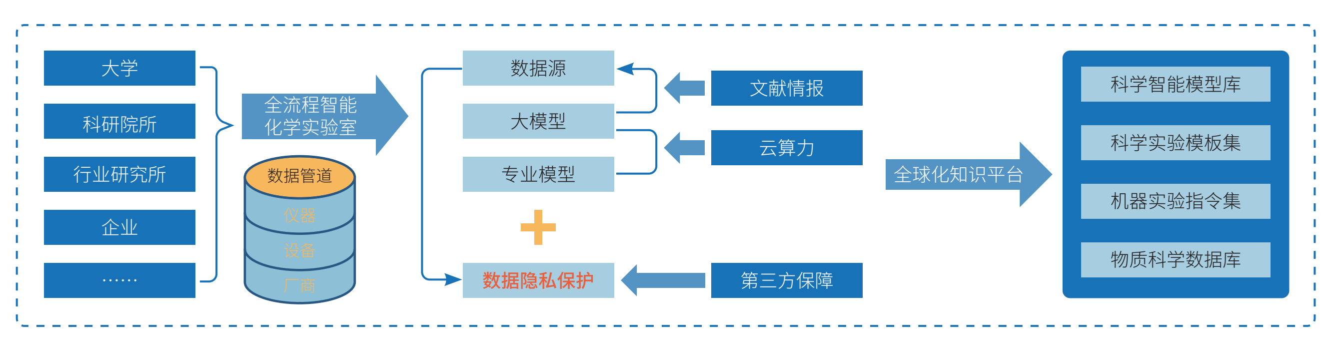


图3　打造精准智能化学的全场景解决方案

Figure 3　Establish full-scenario solution for precision and intelligent chemistry

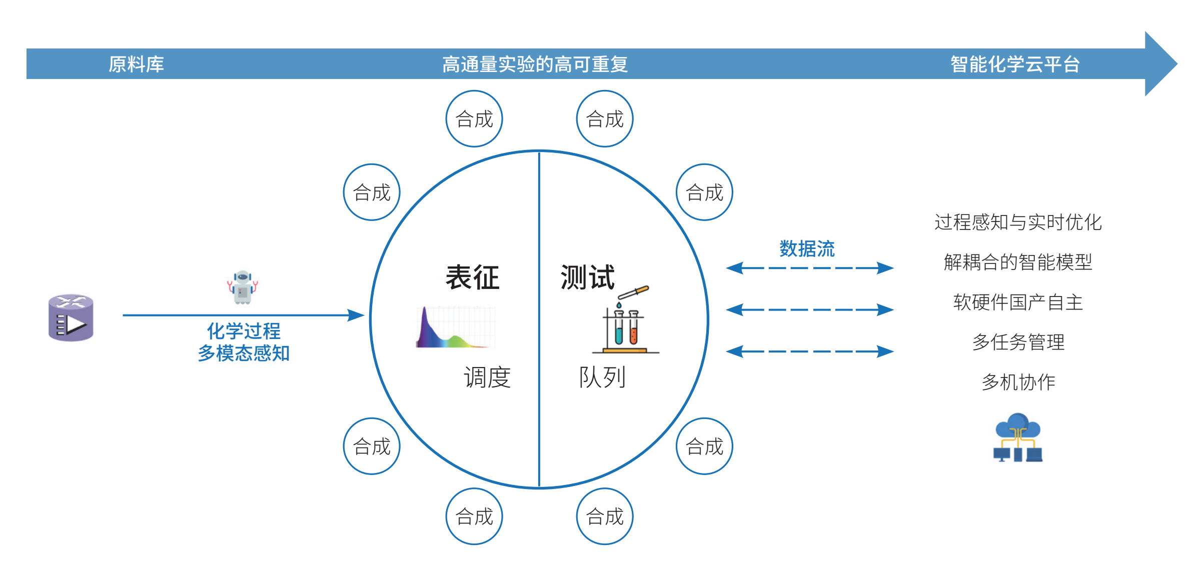


图4　智能化学云平台实现全流程智能控制决策与任务调度

Figure 4　Robotic AI-chemist cloud facility enables intelligent control of decision-making and task scheduling from end-to-end

现智能管理决策，促进不同实验室间操作的高效性和一致性。

4 结语：依托机器化学家云设施变革化学研究范式

以蒸汽机和电机为代表的前两次“碳基”工业革命帮助人类突破了“体力”的限制，以计算机为代表的第三次“硅基”信息技术革命帮助人类突破了“算力”的限制。智能时代已经到来，突破人类“脑力”限制的第四次通用智能工业革命已呼之欲出。呼应时代浪潮，机器化学家云设施的化学科学数据库、科学大模型、机器人平台、智能管理决策系统，将集成解放人类科研人员的“记忆力、体力、算力、脑力”限制，打破科研过程中的知识壁垒、空间限制、学科界限，智能连接科研个体并大幅度拔高其科研能力，全面变革我国化学科学乃至整个物质科学研究范式。

参考文献

1 Li W, Li Z, Zhang H, et al. Efficient catalysts of surface hydrophobic Cu-BTC with coordinatively unsaturated Cu(I) sites for the direct oxidation of methane. PNAS, 2023, 120(10): e2206619120.

2 Dereka B, Yu Q, Lewis N H C, et al. Crossover from hydrogen to chemical bonding. Science, 2021, 371: 160-164.

3 Majerle A, Hadži S, Aupič J, et al. A nanobody toolbox targeting dimeric coiled-coil modules for functionalization of designed protein origami structures. PNAS, 2021, 118(17): e2021899118.

4 Mou T, Pillai H S, Wang S, et al. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. Nature Catalysis, 2023, 6(2): 122-136.

5 Yang J, Huang Y, Qi H, et al. Modulating the strong metal-support interaction of single-atom catalysts via vicinal structure decoration. Nature Communications, 2022, 13(1): 4244.

6 Sun Y, Dai S. High-entropy materials for catalysis: A new frontier. Science Advances, 2021, 7(20): eabg1600.

7 Rai S K, Khanna R, Avni A, et al. Heterotypic electrostatic interactions control complex phase separation of tau and prion into multiphasic condensates and co-aggregates. PNAS, 2023, 120(2): e2216338120.

8 Liang X, Fu H, Crosby A J. Phase-transforming metamaterial with magnetic interactions. PNAS, 2022, 119(1): e2118161119.

9 Li Y M, Yuan J, Ren H, et al. Fine-tuning the micro-environment to optimize the catalytic activity of enzymes immobilized in multivariate metal-organic frameworks. Journal of the American Chemical Society, 2021, 143(37): 15378-15390.
10 Tang T C, An B, Huang Y, et al. Materials design by synthetic biology. Nature Reviews Materials, 2021, 6(4): 332-350.
11 Hirschi S, Ward T R, Meier W P, et al. Synthetic biology: Bottom-up assembly of molecular systems. Chemical Reviews, 2022, 122(21): 16294-16328.
12 Jousset A, Eisenhauer N, Merker M, et al. High functional diversity stimulates diversification in experimental microbial communities. Science Advances, 2016, 2(6): e1600124.
13 Sun Y, Latora V. The evolution of knowledge within and across fields in modern physics. Scientific Reports 2020, 10 (1): 12097.
14 E W N. The Dawning of a new era in applied mathematics. Notices of the American Mathematical Society, 2021, 68(4): 121-130.
15 Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science, 2009, 324: 81-85.
16 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484-489.
17 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596, 583-589.
18 Mehr S H M, Craven M, Leonov A I, et al. A universal system for digitization and automatic execution of the chemical synthesis literature. Science, 2020, 370: 101-108.
19 Burger B, Maffettone P M, Gusev V V, et al. A mobile robotic chemist. Nature, 2020, 583, 237-241.
20 Zhu Q, Zhang F, Huang Y, et al. An all-round AI-chemist with a scientific mind. National Science Review, 2022, 9 (10): nwac190.

Large model-driven, human-computer collaborative robotic AI-chemist cloud facility

CHONG Yuanyuan FENG Shuo WANG Song JIANG Jun*

(Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei 230026, China)

Abstract At present, chemical science is facing unprecedented opportunities and challenges due to the technological changes brought by artificial intelligence. In order to promote the paradigm shift in chemical research, this study proposes the construction plan of the robotic AI-chemist cloud facility. This system realizes a new paradigm of scientific research by collecting multi-channel data to build a database, developing large scientific models enhanced by chemical knowledge, constructing clusters of robotic facilities, and building an intelligent management decision system, which will dramatically improve the efficiency of scientific research and solve scientific problems in terminal applications. This infrastructure is expected to change the paradigm of research and lead to major scientific breakthroughs in chemistry.

Keywords paradigm shift in chemical research, robotic AI-chemist cloud facility, artificial intelligence, automated experiments, human-computer collaboration

崇媛媛 中国科学技术大学博士研究生。从事理论与计算化学、机器学习方向研究，针对化学小数据建立了可解释、可解耦的智能模型，实现了材料的高效筛选。E-mail: yychong@ustc.edu.cn

CHONG Yuanyuan Ph. D. student at University of Science and Technology of China. Her research focuses on theoretical and computational chemistry as well as machine learning. She has developed interpretable and decoupled intelligent models for small chemical datasets, enabling efficient material screening. E-mail: yychong@ustc.edu.cn

江 俊 中国科学技术大学精准智能化学重点实验室副主任。从事理论化学研究，发展融合人工智能与大数据技术的量子化学方法，聚焦于复杂体系内电子运动模拟，研究在多个物理与化学应用领域（能源催化、功能材料、光化学、谱学）中的实际问题。E-mail: jiangj1@ustc.edu.cn

JIANG Jun Associate Director of the Key Laboratory of Precision and Intelligent Chemistry at University of Science and Technology of China. Engaged in theoretical chemistry research, he focuses on developing quantum chemistry methods that integrate AI and big data technologies. His research centers on simulating electronic movements within complex systems, addressing practical issues across multiple physics and chemistry domains such as energy catalysis, functional materials, photochemistry, and spectroscopy. E-mail: jiangj1@ustc.edu.cn

■责任编辑：文彦杰

*Corresponding author